
MASSACHUSETTS
LIBRARIES
INSTITUTE of TECHNOLOGY

High Frequency Data and Volatility in Foreign Exchange Rates

by

Bin Zhou

MIT Sloan School Working Paper 3485-92
September 1992

High Frequency Data and Volatility in Foreign Exchange Rates

by

Bin Zhou

MIT Sloan School Working Paper 3485-92
September 1992

High Frequency Data and Volatility in Foreign Exchange Rates

Bin Zhou
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

September 1992

Abstract: Exchange rates, like many other financial time series, display substantial heteroscedasticity. This poses obstacles in detecting trend and changes. Understanding volatility becomes extremely important in studying financial time series. Unfortunately, estimating volatility from low frequency data, such as daily, weekly, or monthly observations, is very difficult. Recent availability of ultra-high frequency observations, such as tick-by-tick data, to large financial institutions creates a new possibility for analysis of volatile time series. Tick-by-tick data provides us a near continuous observation of the process that gives us potential to study volatility in much detail. However, high frequency data has extremely high negative first order autocorrelation in its return. In this paper, we use tick-by-tick Deutsche Mark and US Dollar (DM/$) exchange rates to explore this new type of data. We propose a model to explain the negative autocorrelation and a volatility estimator using the high frequency data. Daily and hourly volatility of the DM/$ exchange rates are estimated and the behaviors of the volatility are discussed.

Key Words: Financial time series; tick-by-tick data; heteroscedasticity;

1 Introduction

There is considerable literature analyzing the behavior of exchange rates. However, structural exchange rate modeling has not been very successful. By studying monthly data, Meese and Rogoff (1983a,b) have shown that a random walk model fits at least as well as more complicated structural models.

Empirical study (Hsieh 1988) has shown that daily returns are approximately symmetric and leptokurtic (i.e., heavy tailed). The autocorrelations are weak but not independent and identically distributed (iid). One explanation of the heavy tailed distribution is the hypothesis that data is independently distributed as a normal distribution whose mean and variance change over time ([6], [8] and [4]). The argument for changing variance of returns is simple. The amount of "information flow" that cause change in prices is not constant over time. Hence there is no reason to believe that the variance of the price changes is constant over time. Clark (1973) and many others (Mandelbrot and Taylor 1969, Praetz 1972) have argued that observed returns come from a mixture of normal distributions. If the random variable X_t denotes the daily return of the price, the conditional distribution of X_t given information is:

$$X_t|\omega_t \sim N(\mu, f(\omega_t)) \tag{1}$$

where ω_t is all the information available at time t. The quantity ω_t could be the number of transactions (Mandelbrot and Taylor 1969), or trading volume (Clark 1973).

One parametrization of this conditional heteroscedasticity was first studied by Engle (1982) where:

$$f(\omega_t) = \alpha_0 + \sum_{i=1}^{p} \alpha_i (X_{t-i} - \mu)^2, \quad \alpha_0 > 0 \quad \alpha_i \geq 0, \quad i = 1, ..., p \tag{2}$$

It is called the autoregressive conditional heteroscedasticity (ARCH) model since the heteroscedasticity is represented in an autoregressive fashion.

Because the ARCH model exhibits the conditional heteroscedasticity present in financial time series and is mathematically easy to manipulate, it has been used to analyze many financial time series. Previous studies found that the ARCH model provides a close approximation to many financial time series.

Since then many other parametrizations, such as the generalized autoregressive conditional heteroscedasticity (GARCH) model (Bollerslev 1986), have been proposed. They captured some characteristics of the volatility such as volatility clustering. But there also exists problems. The parameters need to be estimated. All these models use historical data. The volatility estimates are often lagged.

The availability of high-frequency data has opened up new possibility in estimating the volatility. Tick-by-tick data provides us with a near continuous observation of the process. It gives us potential to study volatility in great detail. Understanding volatility is the key issue in the conditional heteroscedasticity model (1) and any other financial time series model. This paper explores tick-by-tick data and uses the data to estimate and study the volatility.

2 High-Frequency Data

Because of fast growing computer power, gathering financial data is easier than ever. Data is no longer recorded daily or weekly. Many large institutions began to collect so called *tick-by-tick* exchange rate in the early eighties.

Different from stock market, the foreign exchange market has no geographical location, and no "business-hour" limitations. The deals are negotiated and traded over the telephone. The transaction prices and trading volume are not known to the public. The exchange rates used for most research are the quotes from large data suppliers such as the Reuters, Telerate, or Knight Ridder. Any market maker can submit new quotes to the data suppliers. The quotes then are conveyed to data subscriber's screens. The data suppliers cover the market information worldwide and twenty-four hours a day. The quotes are intended to be used by market participants as a general indication of where exchange rates stands, but does not necessarily represent the actual rate at which transactions are being conducted. It is possible for some participants to manipulate indicative prices occasionally and create a favorable market movement. However, since a bank's reputation and credibility as a market maker emerges from favorable relations with other market participants, it is generally felt that these indicative prices would closely match the true prices experienced in the market. A reader who is unfamil-

Table 1: A Sample of Tick-by-tick Exchange Rates

PUB.	UNIX TIME	BANK	LOC.	RATE
263	632672082	BERGEN BK	OSL	1.6980 -90
WRLD	632672083	SBZX		1.6980/90
263	632672083	COCO	COP	1.6985 -95
263	632672085	AKTIVBANK	VEJ	1.6988 -95
263	632672088	CHEMICAL	N Y	1.6985 -90
263	632672089	CHEMICAL	LDN	1.6987 -92
263	632672091	SWISS BANK	BAS	1.6980 -90
263	632672092	SE BANKEN	MAL	1.6985 -90
263	632672094	MIDLAND BK	LDN	1.6983 -93
WRLD	632672095	DBNY		1.6985/90
263	632672098	SOC GEN	PAR	1.6985 -90

iar with this type of data in the foreign exchange market may want to read Goodhart and Figliuoli's (1991) paper for details. Goodhart and Figliuoli studied minute-by-minute exchange rates (the closing tick of a minute) from the Reuters. They found that the series exhibited (time varying) leptokurtosis, unit roots, and first-order negative correlation. The paper used only three days' data from the Reuters.

In this study, tick-by-tick data for the entire year of 1990 is used. The data is provided by Morgan Guaranty Trust Company (J. P. Morgan). It contains the spot rate quotes from the Reuters and Telerate. To save our computation time, this study concentrates on the Deutsche Mark and US Dollar (DM/$) which is the most active exchange rate in the market. One year of data yields more than two million observations. Each record of data contains following information: data publishers, UNIX time stamp, originator of the data, location, bid and ask prices. A sample of data is listed in Table 1. The rate quoted in the form x.xxxx and 0.0001 is called one basis point. The spread between bid and ask is most often 10 points or less. Since banks are not obligated to trade at the price they quoted, we find quite a few keying errors in the data. To illustrate the problem, a small portion of the raw data is

Table 2: Summary Statistics of Tick-by-tick Returns

n	= 2129364	Med.	= .000000	Skew.	=-0.0399
Mean	=-5.5486e-8	Min.	=-.006621	Kurt.	=10.7544
sd	= 2.3970e-4	Max.	= 0.007515	ρ_1	=-0.4636

plotted in Figure 1. There are quite a few "outliers" in the data. If one could trade at those "outliers", it would have tremendous arbitrage opportunities which one buys or sells the currencies against the jumps. Before the data can be used for statistical analysis, a validation procedure is necessary to remove the "outliers". Our validation procedure removes any sudden jumps with a reversal. The detailed program is listed in appendix II. The stared values in Figure 1 are identified as the "outliers" by our validation program. The validated data are plotted in Figure 2. A sample of removed data has been manually checked. In most cases, the cause of the errors is identified as keying mistake (see Appendix II).

Since the consecutive prices are nonstationary, it is appropriate to study changes in prices. Like many other researches, we prefer to study the compound return which is defined as the difference in the logarithmic value of the bid prices. Table 2 shows the summary statistics of the tick-by-tick returns.

The number of ticks in each minute varies greatly. It ranges from zero to several hundreds. The average return of the tick-by-tick data is negligible in comparison to its standard deviation. Positive and negative moves are equally likely. The returns are skewed slightly to the left. The kurtosis is much higher than three which is the kurtosis of a normal distribution.

Whistler (1988) indicated that kurtosis rises as periodicities become shorter and frequencies become higher by comparing moments of hourly, daily, weekly, and monthly returns. However, Goodhart and Figliuoli (1991) disputed with this finding and found that kurtosis for minute-by-minute return is less than those for hourly and daily returns. To investigate this phenomenon, we calculated the sample kurtosis of return of every n-tick with $n = 1, 2, ..., 1000$ and plotted results in Figure 3. This figure shows that as the frequency increases (n decreases), the sample kurtosis passes three stages: rises, becomes

Figure 1: Original Quotes From the Reuters and Telerate

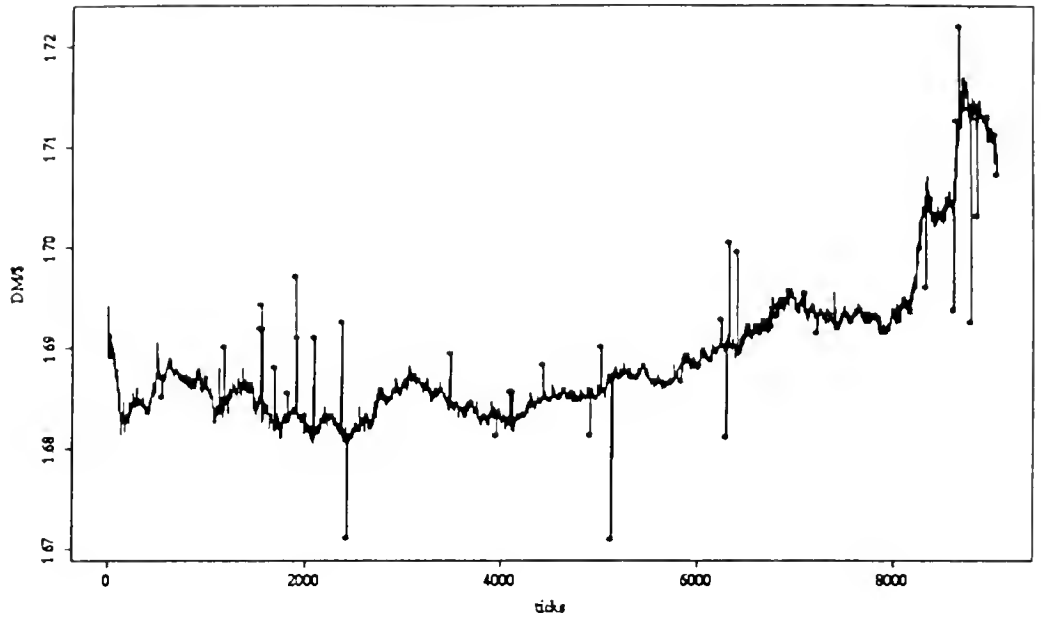


Figure 2: Validated Quotes

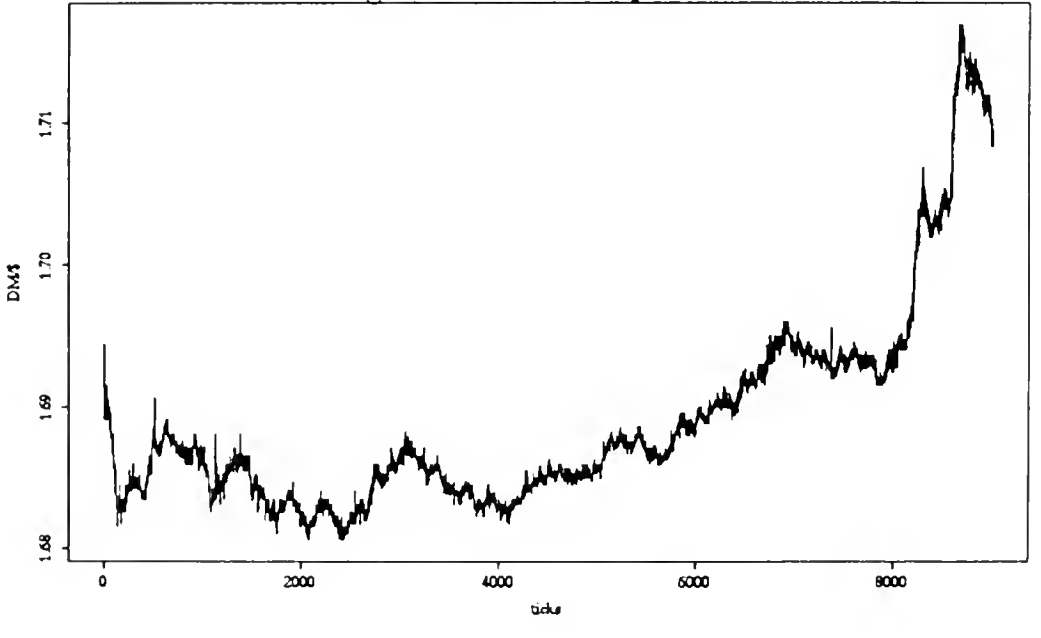
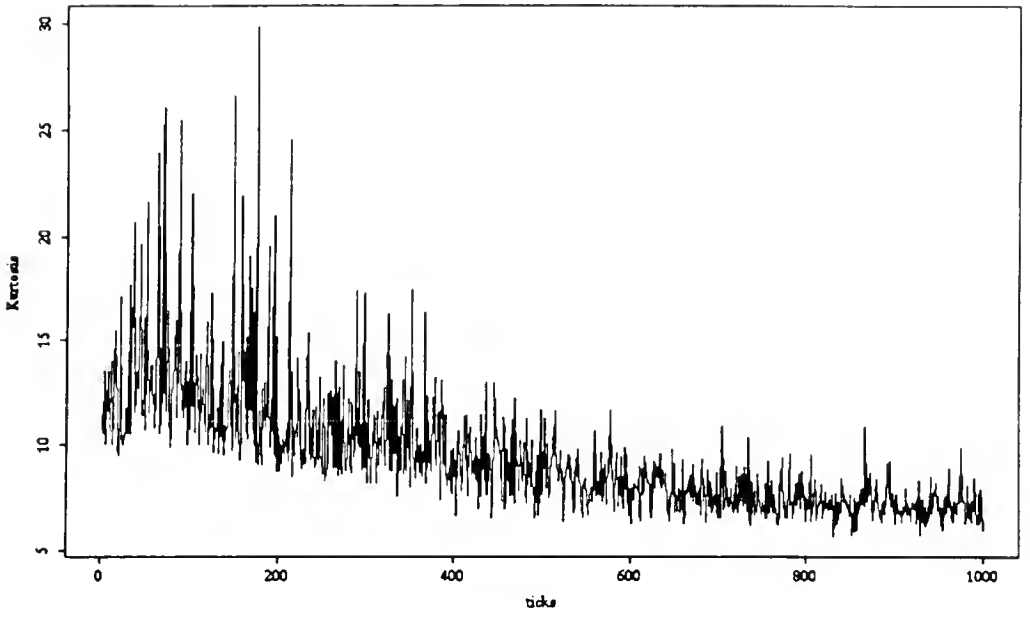


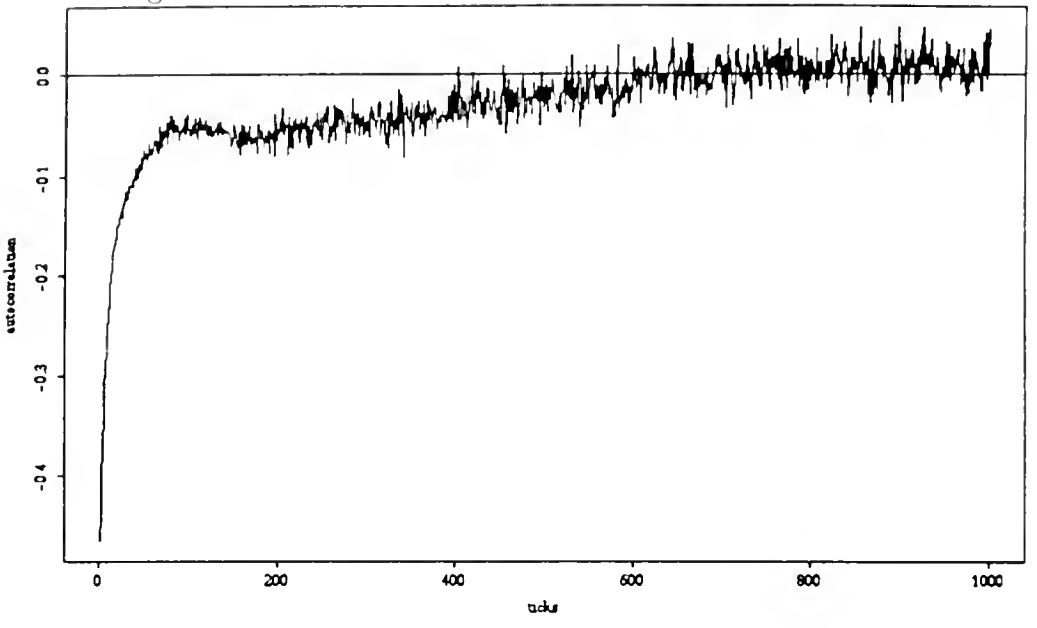
Figure 3: Sample Kurtosis of Return of n-ticks



unstable and then decreases. The last stage, decreasing kurtosis, is due to increasing negative autocorrelation in data (see Figure 4).

Although we expect a slightly negative first lag autocorrelation as it reported in other exchange rate literature, a -47% negative autocorrelation in tick-by-tick return was surprising. To be cautious, we also calculated the autocorrelations in four subgroups according to four quarters. The autocorrelation coefficients are -0.4718, -0.4691, -0.4665 and -0.4632. The negative autocorrelation is consistent. Obviously, high frequency data does not follow a Brownian motion as it is assumed for low frequency data. Our question is if there is any fundamental difference between the high and low frequency data. After further studying the data, we find that the difference in high and low frequency data is level of noises. The noise is negligible in low frequency data, but becomes very significant in high frequency data. The noise may come from many different sources. For example, it could be round off error. All financial data is quoted in finite digits. As we know that round off Brownian motion introduces negative autocorrelation in its return. There are also updating noises in quotes. To be visible in the market, traders keep

Figure 4: First-order Autocorrelations of Return of n-ticks



updating their quotes. The new update is often slightly different from the previous quotes even the market is still. Typographical errors are another source of noise. Summarizing these arguments, we assume following process for the exchange rate:

$$S(t) = d(t) + B(\tau(t)) + \epsilon_t \tag{3}$$

where $S(t)$ is logarithm of the exchange rate, $B(\cdot)$ is the standard Brownian motion, both $d(\cdot)$ and $\tau(\cdot)$ are assumed deterministic functions, $\tau(\cdot)$ has positive increments, and ϵ_t is the mean zero random noise independent to the Brownian motion $B(.)$. The noise, ϵ_t, is combination of several sources which were mentioned above. This extra noise causes most of the negative autocorrelation in high frequency data.

Let $X(s,t) = S(t) - S(s)$, the return in interval $[s,t]$. Then

$$X(s,t) = \mu(s,t) + \sigma(s,t)Z_t + \epsilon_t - \epsilon_s \tag{4}$$

where Z_t is a standard normal random variable, $\sigma^2(s,t) = \tau(t) - \tau(s)$ and $\mu(s,t) = d(t) - d(s)$. The variance of the return is:

$$\mathrm{Var}(X(s,t)) = \sigma^2(s,t) + \eta^2(t) + \eta^2(s) + 2c(s,t)$$

where $\eta^2(t) = \text{Var}(\epsilon_t)$ and $c(s,t) = \text{Cov}(\epsilon_s, \epsilon_t)$. When $|s-t|$ increases, $\sigma^2(s,t)$ increases as well. For large $|s-t|$, noise becomes negligible and $X(s,t)$ behaves like a random walk. When $|s-t|$ decreases to near zero, $\sigma^2(s,t)$ diminishes. The return, $X(s,t)$, is the difference of two noises. The sample first order autocorrelation of such series is about $-.5$. When we study high-frequency data, the noise is no longer negligible. An autocorrelation of -47% for the DM/$ exchange rate indicates that level of noises is very high in tick-by-tick data.

There are several difficulties in analyzing the process (3). One of the difficulties is lack of information about $\tau(t)$ which we call it the *cumulative volatility*. $\sigma^2(t-\delta,t) = \tau(t) - \tau(t-\delta)$ is called the δ-increment of volatility or simply *δ-volatility*. Next section, we devote our attention to estimate the volatility for any increment.

3 Volatility Estimation

In this section we concentrate on estimating the volatility of a given time $[0,n]$, $\tau(n) - \tau(0)$. The function $\tau(t)$ can be estimated increment by increment. We first derive an optimal estimator of the volatility based on assumption of constant variance and zero mean. Then we generalize the results to general process (3). Proofs of the theorems are listed in appendix I.

Theorem 1 *Assume that $\{S(t), t = 0,1,...,n\}$ is a series of observations from the process*

$$S(t) = B(\tau(t)) + \epsilon_t \tag{5}$$

where $\epsilon_t, t = 1,..,n$, are independent and identically distributed with normal distribution and $\tau(t) = \sigma^2 t + b$. Let $X_t = S(t) - S(t-1)$. Then the maximum likelihood estimator of σ^2 is

$$\hat{\sigma^2}_{MLE} = (1/n)\sum_{i=1}^{n}(X_i^2 + 2X_iX_{i-1}\frac{\rho}{\rho'})\frac{(1-\rho\rho')}{(1-\rho^2)} \tag{6}$$

where

$$\rho = \frac{\sum_1^n X_iX_{i-1}}{\sum_1^n X_{i-1}^2} \qquad \text{and} \qquad \rho' = \frac{\sum_1^n X_iX_{i-1}}{\sum_1^n X_i^2}$$

This MLE is not unbiased. However, we noticed that ρ and ρ' are very close. An unbiased estimator can be obtained by eliminating the factors $\frac{(1-\rho\rho')}{(1-\rho^2)}$ and $\frac{\rho}{\rho'}$:

$$\hat{\sigma^2}_U = (1/n)\sum_1^n (X_i^2 + 2X_iX_{i-1}). \tag{7}$$

Theorem 2 *Under the assumptions of Theorem 1, the mean and variance of the estimator (7) are:*

$$\mathbf{E}\hat{\sigma^2}_U = \sigma^2 \tag{8}$$

and

$$\mathbf{Var}(\hat{\sigma^2}_U) = (1/n)\sigma^4(6 + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{\sigma^4}) + 4\eta^4/n^2. \tag{9}$$

From (9), we find that variance of $\hat{\sigma^2}_U$ can be optimized by properly adjusting the variance ratio η^2/σ^2. Since aggregation increases the variance σ^2, we apply estimator (7) to $X_{i,k} = S(i) - S(i-k), i = k, 2k, ..., n$ where we assume that n is multiple of k. Let

$$\hat{\sigma^2}_{U,k} = \frac{1}{n}\sum_{i=k,2k,...,n} (X_{i,k}^2 + 2X_{i,k}X_{i-k,k}). \tag{10}$$

Estimator (10) is unbiased and the variance is given in following theorem:

Theorem 3 *Under the assumptions of theorem 1*

$$\mathbf{Var}(\hat{\sigma^2}_{U,k}) = (1/n)\sigma^4(6k + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{k\sigma^4}) + 4\eta^4/n^2 \tag{11}$$

The variance is minimized at $k = [\frac{2}{\sqrt{3}}\frac{\eta^2}{\sigma^2}]$ *or* $k = [\frac{2}{\sqrt{3}}\frac{\eta^2}{\sigma^2}] + 1$, *where* $[x]$ *rounds* x *down to the next integer.*

Since the noise ϵ_t is independent, the variance of the estimator can be further reduced by averaging the estimator (7) at different start points:

$$\hat{\sigma^2} = \frac{1}{kn}\sum_{i=1}^n (X_{i,k}^2 + 2X_{i-k,k}X_{i-2k,k}) \tag{12}$$

In fact, it can be proved that:

$$\mathbf{Var}(\hat{\sigma^2}) \leq \frac{1}{n}\sigma^4(6k + 16\frac{\eta^2}{k\sigma^2} + 8\frac{\eta^4}{k^2\sigma^4}) + 4\eta^4/n^2 \tag{13}$$

The above three theorems assumed iid noises and constant variances. However the estimator can also be used in a more general case. Suppose that we have observations $\{S(t_i), i = -2k, -2k+1, ..., n\}$ from process (3). Then the volatility $\tau(t_n) - \tau(t_0)$ can be estimated by:

$$V(t_0, t_n) = \frac{1}{k}\sum_{i=1}^{n}(X_{i,k}^2 + 2X_{i-k,k}X_{i-2k,k}) \tag{14}$$

Theorem 4 *Assume $Cov(\epsilon_i, \epsilon_{i-k})$=0 for all i. Then*

$$\begin{aligned}
\mathbf{E}V(t_0, t_n) &= \tau(t_n) - \tau(t_0) \\
&\quad + \sum_{i=1}^{k-1}(i/k)[\sigma^2(t_{i-1-k}, t_{i-k}) - \sigma^2(t_{n-i+1}, t_{n-i})] \\
&\quad + (1/k)\sum_{i=0}^{k-1}[\eta^2(t_{i-k}) - \eta^2(t_{n-i})] \\
&\quad + (1/k)\sum_{i=0}^{n}[\mu^2(t_i, t_{i-k}) + 2\mu(t_i, t_{i-k})\mu(t_{i-k}, t_{i-2k})]
\end{aligned} \tag{15}$$

where $\eta^2(t) = Var(\epsilon_t)$.

Only assumption we made in this theorem is the uncorrelated noises. $\tau(t)$ can be any increasing function and ϵ_t is not necessary stationary. Since

$$\sum_{i=k}^{n}[\mu^2(t_i, t_{i-k}) + 2\mu(t_i, t_{i-k})\mu(t_{i-k}, t_{i-2k})] \leq 3\max\{|\mu(t_i, t_{i-k})|\}[d(n) - d(0)]$$

the last term in (15) is negligible in high frequency data if the drift $d(t)$ is smooth. Therefore, for large n, the estimator (14) is approximately unbiased if no jumps occurred in time interval $[t_0, t_n]$. This estimator is also easy to be updated when new data becomes available. It will allow us to estimate the volatility $\tau(t)$ dynamically.

4 Estimating Volatility of Exchange Rates

In this section, we apply the volatility estimator (14) to DM/$ exchange rates. The data has one discontinuity which was in the week of August 13 when the data base was shutdown due to a power outage accident in lower Manhattan area. The return of that week is set to be zero and so does the volatility.

To choose the parameters k, we estimated the variance ratio η^2/σ^2 to be approximate 6. Minimizing upper bond of variance (13), we have $k = 6$. Using all available tick-by-tick data, we estimate the volatility of DM/$ exchange rate in entire 1990. The estimate is .010349. To verify this estimate, we compare it to the estimate under the Brownian motion assumption. If the data had no noise and followed a Brownian motion, the quadratic variation

$$Q_k = \sum_{i=k}^{(n/k)} X(t_{ik-k}, t_{ik})^2$$

would be a standard estimator of the volatility. When the quadratic variation is used on data with noise, it overestimates the volatility. The bias decreases along with the sample frequency. The expectation of the quadratic variation is:

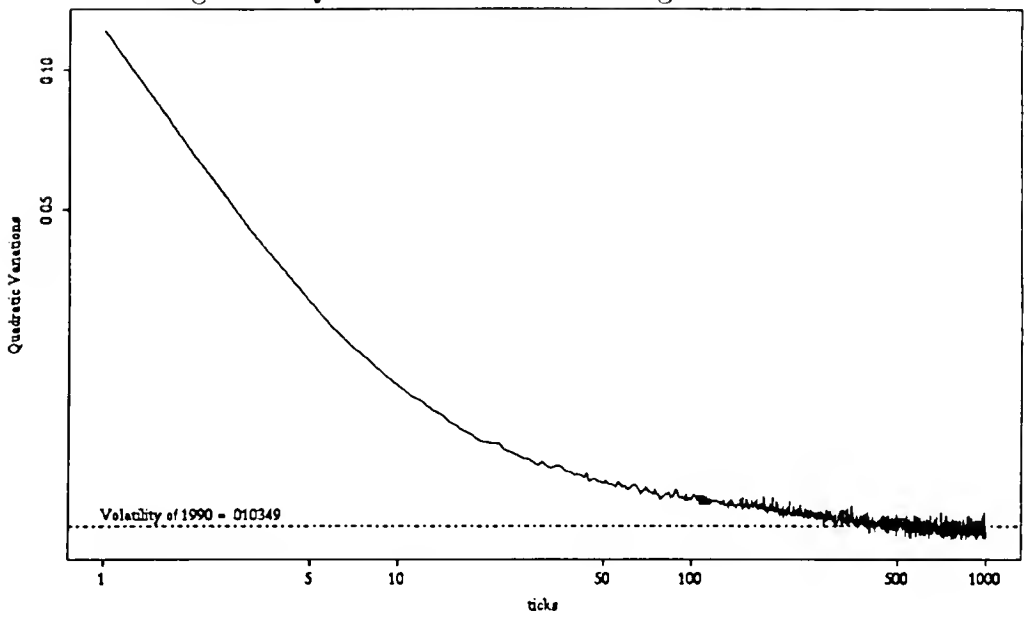
$$\begin{aligned} \mathrm{E}Q_k &= \sum_{i=1}^{(n/k)} [\tau(t_{ik}) - \tau(t_{ik-k}) + \eta^2_{t_{ik}} + \eta^2_{t_{ik-k}} + c(t_{ik}, t_{ik-k}) + \mu^2(t_{ik}, t_{ik-k})] \\ &= \tau(t_n) - \tau(t_0) + \sum_{i=1}^{(n/k)} [\eta^2_{t_{ik}} + \eta^2_{t_{ik-k}} + c(t_{ik}, t_{ik-k}) + \mu^2(t_{ik}, t_{ik-k})] \quad (16) \end{aligned}$$

where $c(t_{ik}, t_{ik-k}) = \mathrm{Cov}(\epsilon_{t_{ik}}, \epsilon_{t_{ik-k}})$. When ϵ_t's are uncorrelated and μ's are negligible,

$$\mathrm{E}Q_k = \approx \tau(t_n) - \tau(t_0) + 2 \sum_{i=1}^{(n/k)} \eta^2_{t_{ik}} \quad (17)$$

which decreases as k increases. We plotted Q_k against k in Figure 5. Both axes have a logarithmic scale. When the frequency is low, the quadratic variation is about the same as our estimate except for a high variation, which is caused by small sample size. In high frequency, the bias is tremendous. When $k = 1$, the quadratic variation is about thirteen times the size of our

Figure 5: Quadratic Variations Using n-tick Returns

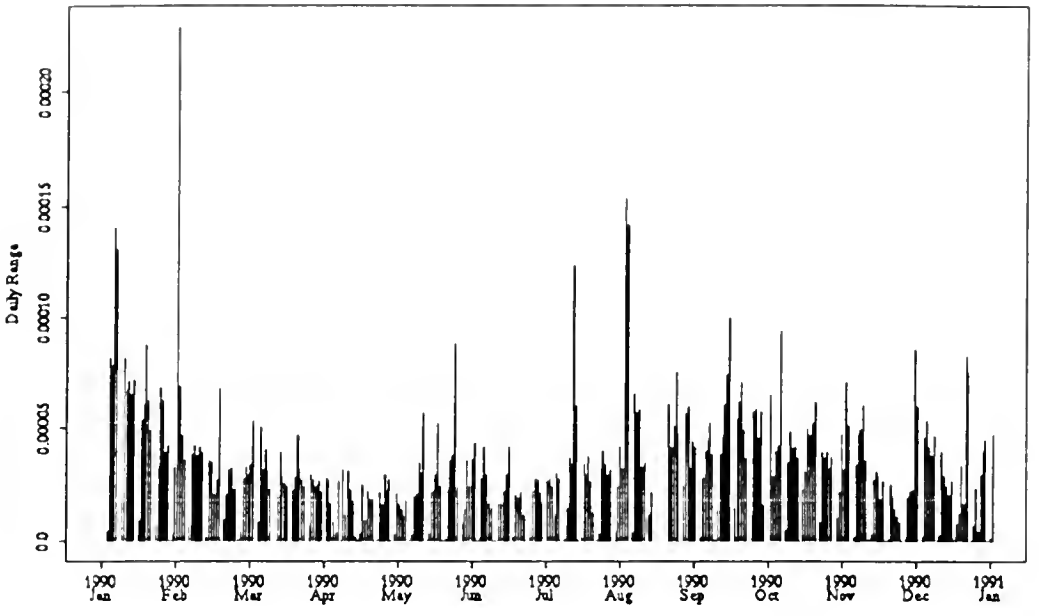


estimate. Therefore, from (17), the total variance of the noise is about six times the total volatility which confirmed our early estimate of the ratio.

To estimate daily volatilities, we need to define the start and the end of a day since the foreign exchange market is a twenty four hour international market. We choose 24 hours from 0:00 Greenwich Mean Time (GMT) as a day because that 0:00 GMT is 9:00am Tokyo time and 24:00 GMT is 7:00pm New York time. This twenty-four hour period covers most activities of the world market. Average weekday daily ticks is more than 7,000. Quotes in weekends or holiday is much less. For small n, volatility estimate from (14) could be negative. In such a case, we let the estimate to zero to avoid negative volatility. The daily volatility estimates of DM/$ are plotted in Figure 6.

The six largest volatilities were on January 4, 5, 30, July 12 and August 2, 3, 1990. On January 4 and 5, the German central bank surprisingly intervened the foreign exchange market and pushed the dollar lower. On January 30, a wide market followed a rumor that Mr. Gorbachev was considering resigning as secretary of the former Soviet Communist Party. On July 12, the dollar tumbled because possible lower interest rate by the US

Figure 6: Daily Volatility Estimates of 1990 DM/$



Federal Reserve. On August 2 and 3, the Dollar had another wild ride as the news of Iraq's invasion of Kuwait spread around the world. However, the large volatility does not always have large price change. The daily change for above six days are -0.0537, 0.0162, 0.0212, -0.0178, 0.0058 and -0.0074 respectively. On August 2 and 3, the exchange rate only changed 58 and 74 points which are about the average.

When we study low frequency data like daily price, noise becomes negligible and the price change, X_i, approximately normal distributed with mean zero and variance σ_i^2, or rescaled return $Y_i = X_i/\sigma_i$ is a standard normal random variable. Therefore we can test our model assumption (3) by testing the normality of rescaled return Y_i. Excluding zero volatility estimates (all on Saturdays), we plot Q-Q normal plots for both return X_i and rescaled return Y_i in Figure 7. The basic statistics of both X_i and Y_i are shown in Table 3. The 95% confidence intervals of the moments of Y_i are given in the parentheses. Table 3 also shows the Kolmogorov-Smirnov goodness-of-fit test for normality. Comparing the statistics in column X_i and column Y_i, we conclude that Y_i is much closer to having a normal distribution. It indicates

Figure 7: Q-Q Plot of Daily Returns

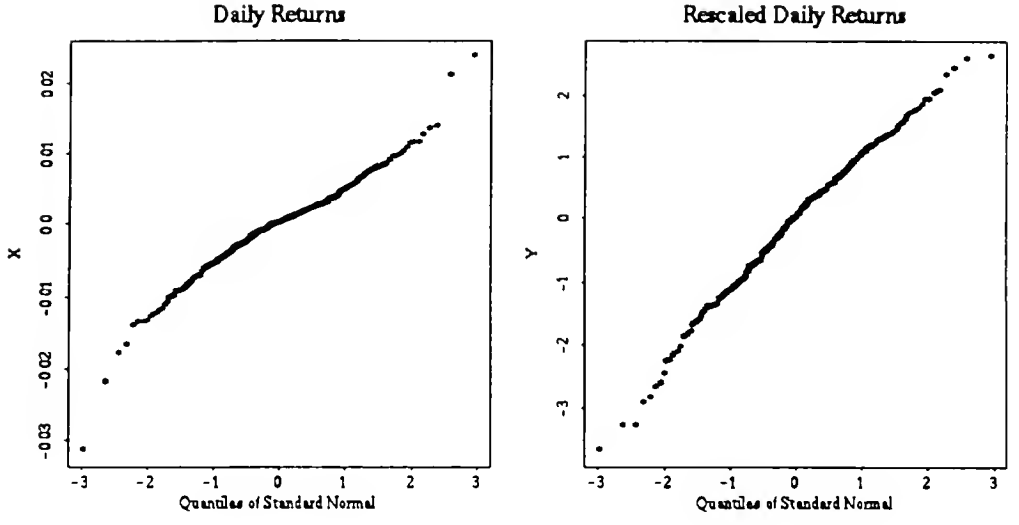


Table 3: Basic Statistics of Daily Returns

	X	Y=X/σ	95% C.I.
n	320	320	
Mean	-0.00031	-0.060	(-0.18, 0.06)
Var	3.62e-5	1.171	(1.01, 1.38)
Skew.	-0.42933	-0.329	(-0.60,-0.06)
Kurt.	6.24585	3.199	(2.65, 3.75)
KS-Test	1.275(p=.00)	0.788	(p=.13)

Table 4: Average Daily Volatility

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Ave. Vol.	7.54e-6	3.25e-5	4.13e-5	3.63e-5	4.45e-5	4.12e-5	7.07e-8
SE	1.11e-6	2.15e-6	4.42e-6	2.43e-6	4.39e-6	3.92e-6	3.19e-8
n	50	51	50	51	51	51	37

that the volatility estimates are reasonably good and the process (3) well describes the high frequency observations of the exchange rates.

To find out if there are any calendar effects in daily volatilities, we calculated seven average daily volatilities. The results are listed in Table 4 as well as plotted in Figure 8. The volatility is low on Monday and high on Thursday and Friday. However, F-statistic testing for equal means shows that the difference is not significant during weekdays.

We repeat the above procedure to estimate hourly volatilities. Total of 6354 hourly volatilities were estimated in 1990. The average number of ticks in an hour is about 335. Using all weekday hourly volatility estimates, we calculate average hourly volatilities which are given in Figure 9. The figure shows that the volatility is high when both the USA and Europe markets are open. A very low volatility around 10:00pm EST(12:00 noon Tokyo time) is due to the lunch hour in Tokyo.

The Q-Q normal plots of hourly return X_i and rescaled hourly return $Y_i = X_i/\sigma_i$ for all $\sigma^2 > 5e-7$ are given in Figure 10. The basic statistics of both X_i and Y_i are shown in Table 5. Sample kurtosis of rescaled return is about 3. The Q-Q plot of rescaled return is almost straight. However, the normality is rejected by the KS-test. However, the normality is not rejected by Shapiro-Wilk test which is considered as more powerful test for the normality. Since Y_i is not a simulated data. It can't be exactly normal. It is not surprising if it is reject the normality by some test. Compare to X_i, Y_i is much closer to having an iid normal distribution. This implies that we have reasonably good estimates of the volatilities.

Figure 8: Average Daily Volatilities

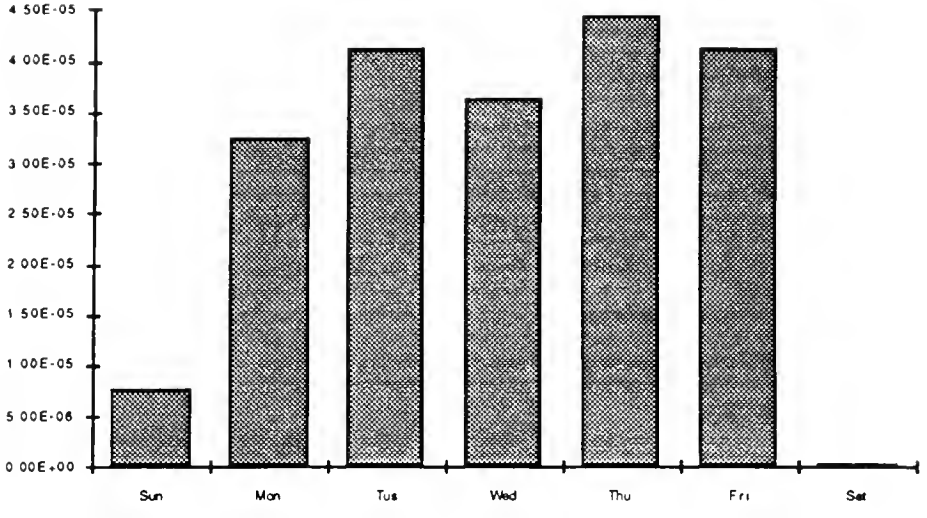


Figure 9: Average Hourly Volatilities

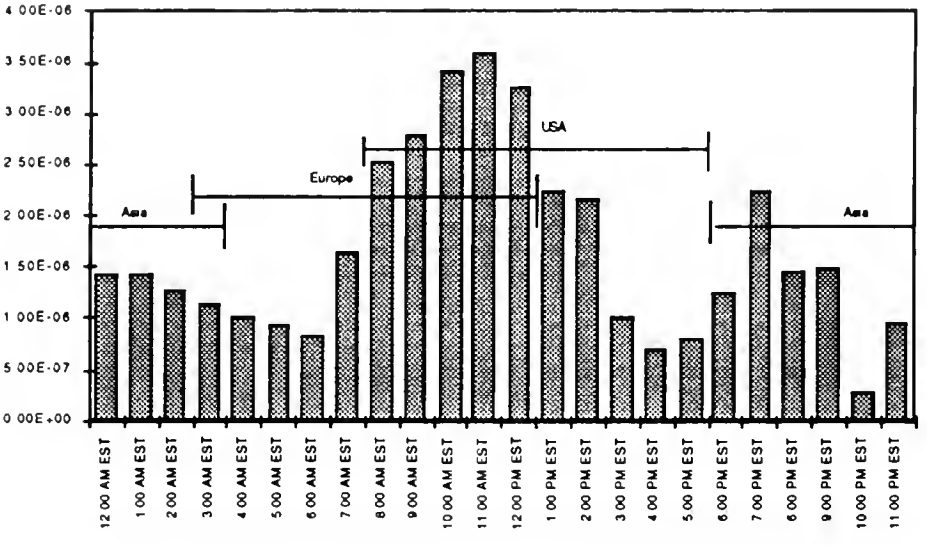


Figure 10: Q-Q Plot of Hourly Returns

Table 5: Basic Statistics of Hourly Returns

	X	Y=X/σ	95% C.I.
n	4377	4377	
Mean	-4.333e-5	-0.0064	(-0.035, 0.022))
var	2.276e-6	0.9127	(0.876, 0.952)
Skew.	-0.255	0.0384	(-0.034, 0.111)
Kurt.	8.32155	2.9501	(2.804, 3.095)
KS-Test	3.847(p=.00)	1.604	(p=.00)

5 Conclusion

High frequency data can be described as a Brownian motion with a noise. This noise brings a strong first lag negative autocorrelation in high frequency data. The autocorrelation decreases as frequency decreases since the role of the noise reduces. High frequency data can be used to estimate the volatility in high frequency in reasonable precision. Different from other volatility estimator, our volatility estimator mainly uses the data within the period we are interested instead of historical data. This allows us to capture the market volatility quickly without delay. The estimate is nearly unbiased when price has no jump. Since trading volume and volatility are highly correlated, this type of volatility estimation is more important in exchange market than in other market because of unknown trading volume.

Besides of many other applications of the volatility, we are more interested in modeling and forecasting the time series. Volatility can be used to rescale the return to address the heteroscadesticity like the ARCH model. We also can analyze the data in volatility scale in stead of calendar time scale to eliminate the heteroscadesticity. We will discuss more of these issues in our future papers.

Appendix I: Proof of Theorems

- Proof of Theorem 1:

  Under assumptions of the theorem,

$$X_t = \sigma Z_t + \epsilon_t - \epsilon_{t-1}$$

  is a normal random variable with mean zero and variance $v^2 = \sigma^2 + 2\eta^2$, where η^2 is the variance of ϵ_t. X_t has the first lag autocorrelation $\rho = -\eta^2/v^2$. The likelihood function of X_t is

$$L(s^2, \rho; X_1, ..., X_n) = f(X_n|X_{n-1})f(X_{n-1}|X_{n-2})...f(X_1|X_0)$$

  Notice that $X_t|X_{t-1}$ is also a normal variable with mean ρX_{t-1} and variance $s^2 = v^2(1-\rho^2)$, we have

$$L(s^2, \rho; X_1, ..., X_n) = (2\pi s^2)^{-n/2} exp(-\sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})^2}{2s^2}).$$

  The log-likelihood function is

$$\ell(s^2, \rho; X_1, ..., X_n|X_0) = -\frac{n}{2} log(2\pi s^2) - \sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})^2}{2s^2}$$

  The partial derivative of the likelihood function with respect to ρ is

$$\frac{\partial \ell}{\partial \rho} = \sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})X_{t-1}}{s^2}$$

  Setting the derivative equal to zero and solve for ρ, we have

$$\hat{\rho} = \frac{\sum_{t=1}^{n} X_t X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2}.$$

  Similarly, for s^2, we have

$$\frac{\partial \ell}{\partial s^2} = -\frac{n}{2s^2} + \sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})^2}{2s^4} = 0$$

or

$$\begin{aligned}\hat{s^2} &= \frac{1}{n}\sum_{t=1}^{n}(X_t - \rho X_{t-1})^2 \\ &= \frac{1}{n}(\sum_{t=1}^{n}X_t^2 + \rho^2\sum_{t=1}^{n}X_{t-1}^2 - 2\rho\sum_{t=1}^{n}X_tX_{t-1})\end{aligned}$$

Substituting ρ by $\hat{\rho}$ in above formula, we have

$$\begin{aligned}\hat{s^2} &= \frac{1}{n}(\sum_{t=1}^{n}X_t^2 - \hat{\rho}^2\sum_{t=1}^{n}X_{t-1}^2) \\ &= \frac{1}{n}\sum_{t=1}^{n}X_t^2(1 - \hat{\rho}\hat{\rho}')\end{aligned}$$

where

$$\hat{\rho}' = \frac{\sum_{t=1}^{n}X_tX_{t-1}}{\sum_{t=1}^{n}X_t^2}.$$

It is easy to show that

$$\sigma^2 = v^2(1+2\rho) = s^2(1+2\rho)/(1-\rho^2)$$

Therefore, the maximum likelihood estimator of σ^2 is

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n}\sum_{t=1}^{n}X_t^2(1+2\hat{\rho})\frac{(1-\hat{\rho}\hat{\rho}')}{(1-\hat{\rho}^2)} \\ &= \frac{1}{n}\sum_{t=1}^{n}(X_t^2 + 2X_tX_{t-1}\frac{\hat{\rho}}{\hat{\rho}'})\frac{(1-\hat{\rho}\hat{\rho}')}{(1-\hat{\rho}^2)}\end{aligned}$$

- Proof of Theorem 2:

$$\begin{aligned}\mathbf{E}(1/n)\sum_{1}^{n}(X_t^2 + 2X_tX_{t-1}) &= (1/n)\sum_{1}^{n}(\mathbf{Var}(X_t) + 2\mathbf{E}X_tX_{t-1}) \\ &= (1/n)\sum_{1}^{n}(\sigma^2 + 2\eta^2 - 2\eta^2) = \sigma^2.\end{aligned}$$

and

$$\mathbf{Var}\hat{\sigma^2}_U = (1/n)^2\mathbf{Var}\sum_{1}^{n}(X_t^2 + 2X_tX_{t-1})$$

$$
\begin{aligned}
&= (1/n)^2 \mathbf{Var}[\sum_{1}^{n}(\sigma^2 Z_t^2 + 2\sigma Z_t\epsilon_t + 2\sigma^2 Z_t Z_{t-1} - 2\sigma Z_t\epsilon_{t-2} \\
&\quad +2\sigma Z_{t-1}\epsilon_t - 2\sigma Z_{t-1}\epsilon_{t-1} - 2\epsilon_t\epsilon_{t-2} + 2\epsilon_{t-1}\epsilon_{t-2}) + \epsilon_n^2 - \epsilon_0^2] \\
&= \frac{1}{n}(6\sigma^4 + 16\sigma^2\eta^2 + 8\eta^4) + \frac{2}{n^2}\mathbf{Var}\epsilon_0^2 \\
&= \frac{1}{n}\sigma^4(6 + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{\sigma^4}) + \frac{4}{n^2}\eta^4.
\end{aligned}
$$

- Proof of Theorem 3:

  Since $\mathbf{Var}X_{i,k} = k\sigma^2$, (9) implies that

$$\mathbf{Var}(k\hat{\sigma^2}_{U,k}) = (k/n)(k\sigma^2)^2(6 + 16\frac{\eta^2}{k\sigma^2} + 8\frac{\eta^4}{k^2\sigma^4}) + 4\eta^4/(n/k)^2$$

  or

$$\mathbf{Var}(\hat{\sigma^2}_{U,k}) = (1/n)(\sigma^2)^2(6k + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{k\sigma^4}) + 4\eta^4/n^2$$

  The variance reaches minimum when

$$k = [\frac{2}{3}\frac{\eta^2}{\sigma^2}] \quad \text{or} \quad k = [\frac{2}{3}\frac{\eta^2}{\sigma^2}] + 1$$

- Proof of Theorem 4:

$$
\begin{aligned}
EV(t_0, t_n) &= (1/k)\sum_{0}^{n}[\sigma^2(t_{i-k}, t_i) + \eta^2(t_i) - \eta^2(t_{i-k}) \\
&\quad +\mu^2(t_i, t_{i-k}) + 2\mu(t_i, t_{i-k})\mu(t_{i-k}, t_{i-2k})] \\
&= [\tau(t_n) - \tau(t_0)] \\
&\quad + \sum_{i=1}^{k-1}(i/k)[\sigma^2(t_{i-1-k}, t_{i-k}) - \sigma^2(t_{n-i+1}, t_{n-i})] \\
&\quad +(1/k)\sum_{i=0}^{k}[\eta^2(t_{i-k}) - \eta^2(t_{n-i})] \\
&\quad +(1/k)\sum_{i=0}^{n}[\mu^2(t_i, t_{i-k}) + 2\mu(t_i, t_{i-k})\mu(t_{i-k}, t_{i-2k})]
\end{aligned}
$$

Appendix II: Validation Program

There are many reasons to have outliers in original data set shown in Figure 1. Most quotes are typed in by human, there are unavoidable keying errors. Most outliers we found are this type of errors. Outlier also could be caused by large bid and ask spread. In such case, at least one of bid or ask price does not reflect true market price and becomes an outlier. Occasionally, electronic error also makes outliers. Following program is designed to remove above three types of outliers:

A quote is considered as an outlier and removed from the time series if

i) a rate is more than 5 or less than 1, or

ii) bid and ask spread is more than 50 points, or

iii) a rate is above or below than its neighbor prices more than a certain threshold.

For (iii), we carry out two regressions using ten bid prices on each side of the data. If the current bid price is higher or lower than c-point from both regressions, it is considered as an outlier, where c is range from 15 to 30 points dependent on the variances of the neighbor points. Whenever an outlier is detected, we go back ten steps and repeat above procedure.

About 0.73% of data have been removed by this validation program. The first ten removed data due to (iii) are listed in Table 6 with possible explanation of error.

Table 6: The First Ten Outliers Removed From 1990 DM/$

Removed Data		Average(Before)		Average(After)		Reason
Bid	Ask	Bid	Ask	Bid	Ask	
1.6938	1.6943	1.68280	1.68377	1.68222	1.68314	Typo
1.6910	1.6920	1.68214	1.68314	1.68278	1.68385	Typo
1.6831	1.6880	1.68141	1.68231	1.68128	1.68217	Spread
1.6847	1.6854	1.68206	1.68286	1.68231	1.68310	
1.6960	1.6970	1.68287	1.68374	1.68271	1.68366	Typo
1.6895	1.6910	1.68242	1.68329	1.68222	1.68319	Typo
1.6825	1.6832	1.68086	1.68186	1.68100	1.68197	
1.6900	1.6910	1.68021	1.68105	1.68051	1.68146	Typo
1.6910	1.6925	1.68101	1.68195	1.68082	1.68179	Typo
1.6700	1.6710	1.68043	1.68125	1.67998	1.68085	Typo

References

[1] Baillie, Richard T. and Tim Bollerslev (1989), "Intra-day and inter-market volatility in exchange rates." *Review of Economic Studies* **58**, 565-585.

[2] Bollerslev, T (1986), "Generalised autoregressive conditional heteroskedasticity." *Journal of Econometrics,* **31**, 307-28. Calderon-Rossel, Jorge and Moshe Ben-Horim (1982), "The behavior of foreign exchange rates, " *J. of Inter. Business Studies,* **13**, 99-111.

[3] Clark, P. K. (1973), "A subordinate stochastic process model with finite variance for speculative price." *Econometrica,* **41**, 135-155.

[4] Diebold, Francis X. (1988), *Empirical modeling of exchange rate dynamics,* Springer-Verlag, New York.

[5] Engle, R.F. (1982), "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation." *Econometrica,* **50**, 987-1008.

[6] Friedman, Daniel and Stoddard Vandersteel (1982), "Short-run fluctuation in foreign exchange rates." *J. Intern. Econ.*, **13**, 171-186.

[7] Goodhart, C.A.E. and L. Figliuoli (1991), "Every minute counts in finacial markets." *J. of Inter. Money and Finance*, **10**, 23-52.

[8] Hsieh, David A. (1988), "The statistical properties of daily foreign exchange rates: 1974-1983." *J. of Inter. Econ.*, **24**, 129-145.

[9] Karatzas, Ioannis and Steven E. Shreve (1988), *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York.

[10] Mandelbrot, B. and H. Taylor (1969), "On the distribution of stock price differences." *Operations Research*, **15**, 1057-1062.

[11] Meese, R. A. and K. Rogoff (1983a), "Empirical exchange rate models of the seventies: do they fit out of sample?." *J. of Inter. Econ.*, **14**, 3-24.

[12] Meese, R. A. and K. Rogoff (1983b), "The out of sample failure of empirical exchange rate models: sampling error or misspecification?." in J. Frenkel, ed., *Exchange Rates And International Microeconomics*, Chicago: University of Chicago Press.

[13] Praetz, P.D. (1972), "The distribution of share price changes." *Journal of Business*, **45**, 49-55.

[14] Taylar, Stephen (1988), *Modeling financial time series*, John Willey & Sons.